

SOME EXPERIMENTS WITH LARGE-VOCABULARY ISOLATED-WORD SENTENCE RECOGNITION

L.R. Bahl, S.K. Das, P.V. de Souza, F. Jelinek, S. Katz, R.L. Mercer, M.A. Picheny.

IBM Thomas J. Watson Research Center, P.O. Box 218 Yorktown Heights, NY 10598

Abstract

This paper deals with two experiments with a large vocabulary isolated word recognizer. The first compares word error rates for 1) meaningful sentences belonging to actual documents and 2) random word lists from the same vocabulary. The error rate is considerably lower for random word lists. The second experiment investigates the performance of the recognition system on sentences containing words outside the vocabulary of the recognizer. Sentences from a 5000 word vocabulary task are recognized with a recognizer limited to a 2000 word subvocabulary. The error rate is only slightly higher than it would be if recognition of the full 5000 word vocabulary was allowed.

Introduction

In a previous paper [1] we presented results on the performance of a speech recognition system for isolated-word sentences from a 5000 word office correspondence task (OC5000). In this paper we present the results of two new experiments with this system.

Our first experiment compares word error rates for two different types of word sequences from OC2000, a vocabulary consisting of the 2000 most frequent words in the office correspondence task. The first type consists of sentences chosen from actual documents. We refer to these as *meaningful sentences*. The second type consists of a random permutation of the words of the same vocabulary, read in sequences of 10 words each. We refer to these as *random word lists*. We decoded the meaningful sentences both with and without a language model, but, of course, the random word lists were decoded without a language model. The experiment with random word lists is similar to experiments for which results have been reported in many papers on large vocabulary recognition [2,3]. We found that the word error rate for meaningful sentences was much higher than for random word lists. This suggests that results based on random word lists are an unrealistic estimate of the performance of a recognition system on meaningful sentences. We believe that there are three reasons for the higher error rate on meaningful sentences: 1) a higher frequency of short words (*e.g., the, a, in*); 2) a higher frequency of homophones (*e.g., to, too, two*); 3) a faster speaking rate. Our results (see Table 1) also show the importance of a language model for recognition of meaningful sentences.

We estimate that the OC5000 vocabulary will cover 96% of natural correspondence text. Users of the system will, of course, sometimes dictate words outside this vocabulary. To see how serious this problem is, we have decoded meaningful sentences from OC5000 using an OC5000 decoder and also using an OC2000 decoder. Limiting the vocabulary of the decoder to OC2000 has two opposing effects on the error rate. For words in the OC2000 vocabulary, the error rate should decrease because some of the possible confusions have been removed by restricting the vocabulary. For words not in the OC2000 vocabu-

lary, the error rate will, of course, be 100%, and these errors may cause additional errors in the surrounding words (*error propagation*) because of the memory of the language model. Our results (see Table 2) were encouraging: the error rate for the OC2000 decoder was only slightly greater than that for the OC5000 decoder. Furthermore, there was no error propagation.

The Recognition System

The office correspondence database consists of approximately 14,000 letters and memos; about 40% of these documents were solicited from IBM customers; the remainder were extracted from an IBM internal electronic mail system. The text was subjected to intensive automatic and manual editing to eliminate headers, closings, duplicate memos, and nonsense memos used to test the electronic mail system. After editing, the size of the corpus was 900,000 words.

The OC5000 vocabulary is obtained by selecting the 4992 words occurring most frequently in the corpus and merging these with a list of 115 words so that all months of the year, days of the week, numerals, and punctuation marks were included. Some of the sentences fully covered by this 5000 word vocabulary were extracted and set aside as test data and for acoustic training. Sample sentences are:

Thanks for your letter dated April six, nineteen seventy eight.

You did an outstanding job of filling in.

We are confident that it will help us sell the use of the DP marketing program.

The OC2000 vocabulary is similarly obtained by merging the 115 word list with the set of 1992 most frequent words in the corpus. After removing sentences to be used for testing, statistics for the two language models were obtained from the remaining database by counting trigrams, bigrams, and unigrams composed of words in the OC5000 and OC2000 vocabularies.

The language model probabilities $P(w_3 | w_2, w_1)$ are obtained as a linear combination of the trigram, bigram, and unigram relative frequencies, as described in [4]. The perplexity [4] of the OC5000 language model is 90, and that of the OC2000 model is 75. In the experiment with no language model, all words in the vocabulary have the same probability of occurrence. Homophones cannot be distinguished acoustically so when the decoder has no language model, it chooses amongst homophones at random.

Recordings were made on a SONY TC-D5M cassette recorder, in a normal office at the IBM T.J. Watson Research Center, using a Shure SM12 close-talking microphone. The same native male speaker of general American English was used in all experiments (RLM in [1]).

26.5.1

For OC5000, training was performed on 100 meaningful sentences. Training for the OC2000 experiments was on random word lists which contained each OC2000 word once. All speech was monitored during recording sessions and was rerecorded if gross mispronunciations were detected. Recordings were digitized at 20 KHz using a 12-bit A/D converter.

The acoustic processor measures the energy in 31 frequency bands each centisecond as described in reference [5]. This 31-parameter vector is then reduced to a label from an alphabet of size 200 by vector quantization [6]. The 200 prototypes used for vector quantization are obtained from 30,000 centiseconds of training data by an automatic clustering method described in [7].

Each word in the vocabulary has a phonetic baseform which describes its pronunciation. Some words have more than one baseform as, for example, α ([e] and [A]). No phonological rules are used. Each phone in the phonetic baseforms is modelled by a Markov source (hidden Markov chain). The statistical parameters of these models are estimated using the Forward-Backward algorithm [4,8]. The decoder is described in reference [4].

Results

Table 1 gives the results for recognition of 100 meaningful sentences (1297 words) from OC2000 with and without a language model, and for recognition of 2070 words from random word lists (one for each baseform in the OC2000 vocabulary). Comparison of the first two error rates (2.5% vs. 20.3%) indicates the power of the language model. Comparison of the last two error rates (20.3% vs. 8.5%) shows that the recognition of meaningful sentences is considerably more difficult than the recognition of random word lists. As mentioned above, with no language model, the decoder chooses amongst homophones at random. Of the 263 words misrecognized in meaningful sentences, 60 were homophone errors. In the random word lists, 32 of the 175 misrecognized words were homophone errors. In comparison, when the language model was used for meaningful sentences, only one of the errors was a homophone error.

Table 2 gives the results for recognition of 50 meaningful sentences from the OC5000 vocabulary using two different decoders: an OC5000 decoder and an OC2000 decoder. The error rate was only slightly higher with the OC2000 decoder. Note, however, that the OC2000 decoder must make errors on at least 4% of the words (these are OC5000 words not in the OC2000 vocabulary). As the OC5000 decoder is improved, the ratio of the error rates for the OC2000 decoder and the OC5000 decoder will necessarily increase. While we have not yet found any robust criterion for automatically detecting words outside the decoder vocabulary, the fact that they do not cause error propagation is encouraging.

Type	Meaningful Sentences	Meaningful Sentences	Random Word Lists
Language model	OC2000	None	None
Total words spoken	1297	1297	2070
Total errors	32	263	175
% errors	2.5%	20.3%	8.5%

TABLE 1. Three OC2000 Recognition Experiments

Language model	OC5000	OC2000
Total words spoken	866	866
Total errors	60	75
% errors	6.9%	8.7%

TABLE 2. Two OC5000 Recognition Experiments

Acknowledgement

We would like to thank Lance Miller for his assistance in obtaining the office correspondence database.

References

1. L.R. Bahl, A.G. Cole, F. Jelinek, R.L. Mercer, A. Nadas, D. Nahamoo, and M.A. Picheny, "Recognition of Isolated-Word Sentences from a 5000-Word Vocabulary Office Correspondence Task," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE Catalog No. 83CH1841-6, Vol. 3, pp. 1065-1067, Boston (April 1983).
2. A.E. Rosenberg, L.R. Rabiner, and J.G. Wilpon: "Speaker Trained Recognition of Large Vocabularies of Isolated Words," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE Catalog No. 82CH1746-7, Vol. 3, pp. 2018-2021, Paris (May 1982).
3. T. Kaneko and N.R. Dixon: "A Hierarchical Decision Approach to Large Vocabulary Discrete Utterance Recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. ASSP-31, No. 5, pp. 1061-1066, October, 1983.
4. L.R. Bahl, F. Jelinek and R.L. Mercer, "A Maximum Likelihood Approach To Continuous Speech Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. PAMI-5, No.2, March 1983.
5. S.K. Das, "Some Dimensionality Reduction Studies in Continuous Speech Recognition," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE Catalog No. 83CH1841-6, Vol. 1, pp. 292-295, Boston (April 1983).
6. A. Buzo, A.H. Gray Jr., R.M. Gray, and J.D. Markel, "Speech Coding Based Upon Vector Quantization," *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. ASSP-28, pp. 562-574, October, 1980.
7. A. Nadas, L.R. Bahl, R. Bakis, P.S. Cohen, A. Cole, F. Jelinek, B.L. Lewis, and R.L. Mercer, "Continuous Speech Recognition with Automatically Selected Prototypes Using Either Bootstrapping or Clustering," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE Catalog No. 81CH1610-5, pp. 1153-1155, Atlanta, (April 1981).
8. L.E. Baum, "An Inequality and Associated Maximization Technique in Statistical Estimation of Probabilistic Functions of Markov Processes," *Inequalities*, Vol. 3, pages 1-8, (1972).